

Elementos para una Propuesta de Aseguramiento de la Calidad de Datos en Analítica

Pérez, C.¹; Basso, D.^{2,3}; Poggi, E.⁴

¹*Programa de Maestría en Informática. Departamento de Ingeniería e Investigaciones Tecnológicas. Universidad Nacional de la Matanza, Argentina*

²*Departamento de Ingeniería e Investigaciones Tecnológicas. Universidad Nacional de La Matanza, Argentina*

³*Grupo de Estudio en Metodologías de Ingeniería de Software. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional, Argentina*

⁴*Universidad de San Andrés, Argentina
tutocari@gmail.com, {diebasso, eduardopoggi}@yahoo.com.ar*

Resumen

La sinergia entre la analítica, las ciencias de datos y Big Data sustentan las bases para gestionar grandes volúmenes de información y su procesamiento de forma ágil [1], permitiendo descubrir patrones y comportamientos de los datos. La posibilidad de dar a los clientes lo que desean a través de la analítica de datos y sus herramientas, es el propósito para lograr la satisfacción de éstos y el crecimiento de las organizaciones. Dicha evolución contiene una gran falencia que todavía se debe solventar, y es la calidad de los datos contenidos en los repositorios para analítica. Si bien existen modelos, normas y estándares que plantean aspectos específicos de la calidad y gobernanza de datos, ninguno de ellos sugiere cómo evaluar los datos desde la perspectiva de la analítica. De acuerdo a los niveles de madurez que se definieron en el modelo MAMD, el aseguramiento de los datos se encuentra en el nivel predecible y la mejora de la calidad de datos en su último nivel, el de innovación. Por ello, la falta de control de la calidad de los datos puede llevar a que la resultante del proceso analítico no sea idónea y veraz perjudicando el negocio.

En este contexto, el presente trabajo tiene como objetivo definir los elementos para desarrollar un marco de aseguramiento de calidad de datos en analítica, que permita que los reportes y tendencias que se obtienen del proceso resulten beneficiosos para el negocio, aportando datos oportunos y de valor.

1. Introducción

Cada vez más se destaca la importancia de la calidad de los datos como activo organizacional [2]. Siempre se habló de la misma respecto de los productos de software, de los procesos y sus fases o bien del aseguramiento de la calidad de la información en las etapas de recolección o

almacenamiento [3]. El aseguramiento de calidad presupone la existencia de un sistema de control de calidad de las mediciones (Quality Control) y de un sistema de evaluación de la calidad (Quality Assessment) [4]. La ausencia de cualquiera de estos componentes compromete la validez de los resultados analíticos [4].

Teniendo en cuenta que los datos provienen de diversas fuentes, éstos deben cumplir en un alto grado con el conjunto de atributos determinados por modelos afines de calidad. La baja calidad de los datos es un factor que afecta el desempeño de los sistemas de información en las organizaciones, debido a que representa uno de los problemas que no se perciben fácilmente y más graves en los procesos transaccionales [5]. Por otro lado, también se advierte que las denominadas 5Vs (Volumen, Velocidad, Variedad, Veracidad y Valor) provocan muchos inconvenientes a la hora de extraer datos reales y de alta calidad por su magnitud, velocidad de cambio y complejidad [6]. En [6] se plantean los desafíos a los que se enfrentan las organizaciones en la era de Big Data:

- Muchas fuentes y tipos de datos
- Gran volumen de datos
- Alta volatilidad
- Inexistencia de estándares de calidad

La analítica de datos ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades [7], transformando los datos en información valiosa. Conocer, clasificar, filtrar y utilizar la información mediante un análisis exhaustivo de los datos es imprescindible para que se conviertan en verdaderos activos y generadores de negocio [8]. Existen modelos y estándares que describen características y principios que una organización debe considerar al evaluar y mejorar la calidad de sus datos, para lograr optimizar los procesos de gobierno y gestión, y reducir los riesgos relativos a los datos, entre los que se destacan

el modelo MAMD [9], la metodología TDQM (Total Data Quality Methodology) [10] y la norma ISO/IEC 25012 [11]. Sin embargo, ninguno de ellos plantea un marco metodológico con procesos que aseguren la calidad de los datos para el procesamiento en analítica. Tampoco se ha encontrado evidencia documental específica que aborde este tema. En este contexto, se ha detectado la necesidad de identificar y especificar los elementos para proponer un marco integral de aseguramiento de la calidad adecuado, que permita generar datos de valor para todos los niveles de la organización en el contexto de la analítica.

El presente trabajo plantea entonces la definición del problema (sección 2), los elementos y metodología utilizada (sección 3), la solución propuesta (sección 4) y los resultados obtenidos y esperados (sección 5). Finalmente se exponen las conclusiones y futuras líneas de trabajo (sección 6).

2. Definición del Problema

La calidad del conocimiento extraído depende en gran medida de la calidad de los datos y una buena analítica precisa de ello [12]. Obtener datos fiables para el análisis avanzado constituye una preocupación para las organizaciones, dado que la falta de exactitud, completitud y consistencia de la calidad de los datos, dificulta la toma de decisiones [12]. Sin información suficiente, las conclusiones se limitan y el resultado de la analítica pierde precisión y profundidad. Frente a la evolución tecnológica y la búsqueda del fortalecimiento y mejora de la información recolectada, no se cuenta con un marco de trabajo que contenga elementos, metodología y procesos para analítica que aseguren la calidad de los datos a la hora de utilizarlos. Estos aspectos, deberían implicar el análisis, clasificación y filtro de los datos que realmente aporten valor para la toma de decisiones. Para poder realizar esto, se debería contar con un marco adecuado de aseguramiento de calidad. El mismo resultaría imprescindible para demostrar y evaluar de manera transparente, objetiva y documentada la validez de los procedimientos para generar información confiable. Haciendo hincapié en el último de los desafíos mencionados en [6], se busca desarrollar un marco de aseguramiento de calidad de datos orientado específicamente para analítica. Este marco, que actualmente se encuentra en proceso de desarrollo, se orientará exclusivamente a datos tomados de un repositorio para analítica, siendo estos los que se evaluarán y clasificarán. Dentro de los objetivos que se plantean para la definición de este marco, se tiene:

- Proponer un estándar de calidad de datos formado a partir de métricas, modelos existentes y normas, que busquen la mejora de calidad o bien puedan brindar un aporte para la misma.

- Construir procesos que vayan resolviendo cuestiones de forma progresiva para obtener activos informacionales de buena calidad.
- Diseñar un modelo para evaluar analíticamente los datos y devolver información útil y de valor para el negocio.
- Establecer una categoría para grupos de datos.
- Establecer niveles de criticidad para los datos.
- Disminuir la cantidad de datos no oportunos o veraces, para que sólo permanezcan aquellos que cumplan el fin para el que fueron tomados.
- Lograr integridad y disponibilidad de los datos en forma adecuada para garantizar su futura utilización.

Teniendo en cuenta lo enunciado, el interrogante que se plantea es poder definir y especificar los elementos necesarios para estructurar un marco metodológico de aseguramiento de calidad de datos en analítica, donde se incluyan además indicadores y aspectos a tener en cuenta. Estos elementos servirán para posteriormente definir una metodología que incluya los procesos en los que se enmarcarán cada una de sus secciones. La metodología y procesos indicados quedan fuera del alcance del presente trabajo.

3. Elementos y Metodología de Trabajo

Dentro de los modelos, estándares y metodologías definidos para la evaluación de la calidad de los datos, se tiene: el modelo MAMD (Modelo Alarcos de Mejora de Datos) [9], centrado en la gestión y gobierno de datos, que incluye un modelo de referencia de procesos y un modelo de evaluación; la norma ISO/IEC 25012 [11], con características de calidad de datos dependientes de los sistemas y otros inherentes a ellos; y la metodología TDQM (Total Data Quality Methodology) [10, 13], cuyo objetivo es la mejora continua de los productos de información en el contexto organizacional.

Para la definición y especificación de los elementos, metodología y procesos aplicables a los datos para analítica que se incluirán en este marco, se está llevando adelante una investigación de tipo clásica-teórica [14, 15]. En la especificación de los elementos, se incluye también la definición de los niveles resultantes de los datos, es decir, cómo quedan encuadrados los mismos según un nivel de criticidad otorgado luego de accionado el marco propuesto. La definición de estos elementos y su posterior aplicación en la propuesta del marco de aseguramiento de calidad, utilizará los métodos y materiales que se describen a continuación:

A. Métodos a utilizar

Se establecen los siguientes métodos y criterios que se seguirán para el desarrollo del trabajo:

- Revisiones sistemáticas [16, 17], de aspectos cualitativos, generando una síntesis de la evidencia disponible del problema planteado.
- Prototipado evolutivo [18], para construir modelos que vayan resolviendo cuestiones de calidad para analítica de forma progresiva, generando su refinamiento mediante el análisis, evaluación y aceptación de especialistas en analítica de datos.

B. Materiales

El desarrollo de la presente propuesta se basa principalmente en el modelo MAMD para la mejora de datos, la norma ISO 25012, 9000, entre otras, la metodología TDQM y la adaptación de procesos y tareas de normas afines a la calidad. A continuación, se menciona un resumen de los aspectos tenidos en cuenta para cada uno de ellos:

- Estándares y normas ISO, poniendo el foco en aspectos específicos de la calidad de datos como en aquellos puntos de interés para esta propuesta de marco, entre las que se incluyen las normas ISO 9000 [19], ISO 9126 [20], ISO 15939 [21], ISO 14598 [22] e ISO 25012 [11].
- Modelo MAMD, propuesto para la gestión de la calidad de datos, los procesos de gestión de calidad y el gobierno de datos. Dicho modelo se desarrolló basándose en las normas ISO 8000 [23] e ISO/IEC 33000 [24].
- Estudio de formatos y tipos de datos de base de datos transaccionales.
- Dimensiones e indicadores propios de estándares jerárquicos de calidad para Big Data [6].
- Evaluación de métricas de software e indicadores de procesos que sean útiles de ser interpretadas para medir la calidad de datos.
- Las fases identificadas en la metodología TDQM: Definir – Medir – Analizar y Mejorar [10]

C. Metodología de Investigación

Para alcanzar los objetivos mencionados se plantea la siguiente metodología de trabajo:

- Realizar una investigación exploratoria sobre estrategias o metodologías de calidad utilizadas en diferentes ámbitos y que hayan introducido una solución de Big Data al negocio.
- Investigar acerca de referencias en el tratamiento de la calidad de los datos en los últimos años, para determinar el nivel de avance sobre el tema.
- Analizar estándares referentes a la calidad de la información, procesos, software y familias de normas ISO para obtener directivas, controles y pautas a seguir respecto de la calidad de los datos.

- Identificar aspectos específicos para medir el valor del dato de acuerdo a las características de negocio de una organización.
- Estudiar modelos de calidad de software, de calidad de procesos y de calidad de datos, análisis predictivo, métricas e indicadores de calidad.

4. Solución Propuesta

Como estructura general del marco de aseguramiento de calidad se proponen tres secciones: el marco convencional, el marco de contexto y el marco del negocio. Estos corresponden a áreas de inspección donde se irá analizando y evaluando cada grupo de datos que se seleccione de un reservorio para analítica. A través de este marco metodológico se busca organizar las etapas por las que debe pasar el dato para determinar su validez de calidad. Estas secciones, no enunciadas en ningún otro modelo o estándar, se sustentan en los aspectos y buenas prácticas definidos en los estándares descriptos de la sección 3, inciso B, con lo que se consideran de innovación. Cada grupo de datos podrá inspeccionarse en cada una de estas secciones. Del análisis de cada marco se obtendrá un nivel de criticidad para cada dato que lo habilitará para su evaluación en el siguiente marco. Si el dato logra pasar por todas las secciones, se enmarcará dentro de un nivel final de calidad. El nivel final corresponde a aquél con el que los datos llegarán a los analistas, permitiendo determinar si los mismos pueden usarse para un procesamiento analítico y haciendo que el resultado del mismo (expresado como reporte, informe, cuadro o gráfico) otorgue el fin buscado y con resultados veraces y útiles para el negocio. Cabe aclarar que, la medición o corrección del dato evaluado, en caso de no resultar de utilidad, no entra dentro del alcance de los elementos y marco que se propone definir. En la Figura 1 se muestra un modelo con los elementos que conforman la estructura propuesta para el marco de aseguramiento de calidad de datos en analítica. El mismo comienza con el reservorio de analítica y continúa con cada uno de los marcos. A continuación se explica cada uno de estos elementos, con los indicadores definidos y las acciones/tareas a realizar.

4.1. Reservorio para analítica

Los grupos de datos se toman de un reservorio para analítica, los cuales suponen un proceso de calidad propio, acorde a los estándares y procesos de control de cada organización y de las herramientas utilizadas para tal fin. Antes del procesamiento analítico que se les realizará para obtener la respuesta a un objetivo buscado, se les aplica una metodología con los procesos enmarcados en cada una de las secciones definidas.

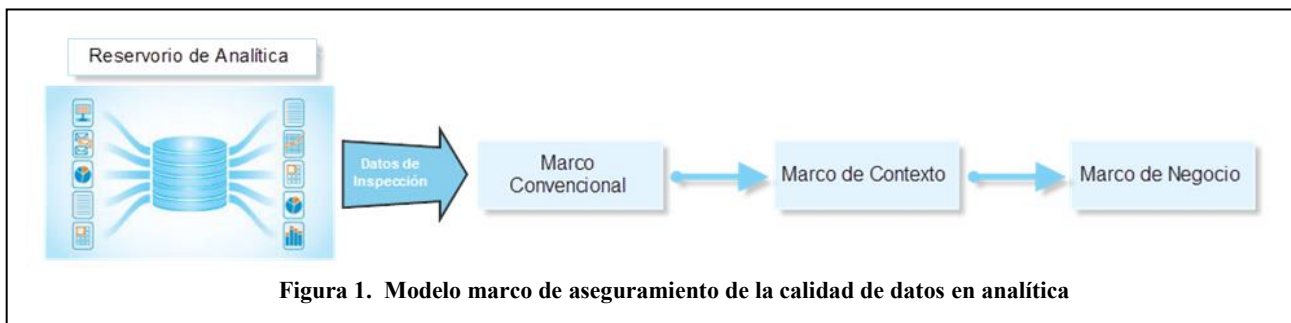


Figura 1. Modelo marco de aseguramiento de la calidad de datos en analítica

Esta información tiene diferentes tipos de datos que pueden provenir de diversas fuentes o proyectos. No obstante, los elementos del marco propuesto sólo se orientan a grupos de datos provenientes de bases de datos transaccionales, excluyendo aquellos provenientes de imágenes, datos de sensores, etc.

4.2. Marco convencional

En una primera instancia, se debe inspeccionar el grupo de datos a nivel físico, evaluando su formato, su tipo y los detalles de calidad que deben cumplir, tal como se especifica en la Tabla 1. En dicha tabla, se sugieren algunas revisiones rápidas y generales a realizar. El término “rápidas” se refiere a que estos chequeos fueron realizados en la base transaccional y al llegar al repositorio de analítica ya pasaron dichos controles. Una vez cumplimentados éstos, al dato resultante se le asigna un nivel de criticidad y una acción a realizar, tal como se indica en la Tabla 2. Aquellos datos que se consideren *de valor* serán los que pasen al “marco de contexto”. Cabe mencionar que, los tipos de datos y formatos válidos deberán cumplir con los estándares de cada organización de acuerdo al servidor de base de datos.

4.3. Marco de Contexto

El análisis de contexto de los datos lleva consigo la tarea de chequear dimensiones de calidad con sus indicadores, verificar su veracidad y valor por medio de análisis estadísticos y observaciones de referencias externas. La calidad de los datos no sólo depende de sus propias características contenidas en el modelo físico, sino también del entorno que utiliza los datos, incluidos los procesos y los usuarios. Sólo aquellos datos que se ajustan, o se sitúan, dentro del contexto analizado y que cumplen con los requisitos definidos, pueden considerarse como calificados o de buena calidad dentro de este marco de contexto.

En las Tablas 3 a 7 se definen las siguientes cinco dimensiones conteniendo diez indicadores de estándares de calidad para Big Data [6]: *Disponibilidad, Usabilidad, Confiabilidad, Pertenencia y Referencia Externa*. Para

cada una ellas, se evalúan los datos resultantes de nivel válido provenientes del marco convencional. Las dimensiones se definieron de acuerdo a su capacidad intrínseca para satisfacer las necesidades propias de los datos, siendo éstas las más aceptadas y utilizadas [6, 11]. Las métricas de evaluación para cada uno de los indicadores quedan fuera del alcance del presente trabajo.

Cabe aclarar que no necesariamente cada dato proveniente del marco anterior debe pasar por cada dimensión o indicadores, ya que puede no corresponder su evaluación en alguno de ellos.

Tabla 1. Análisis del Marco Convencional

Tipo de Dato	Descripción
Dirección	Valor alfanuméricos cortos
Apellidos y Nombres	Valor alfabéticos cortos
Edad	Valor numérico entero corto
Importes	Valor numérico grande con decimales
Números e Identificadores (boletas, documentos, expedientes, id)	Valor alfanumérico grande sin decimales.
Geo posiciones	Valor con coordenadas en formato válido. - grados, minutos y segundos (DMS): 41°24'12.2"N 2°10'26.5"E - grados y minutos decimales (DMM): 41 24.2028, 2 10.4418 - Grados decimales (DD): 41.40338, 2.17403.
Fechas	Valor tipo date con formato utilizado en la organización.
Cuentas bancarias	Datos alfanuméricos por lo general de formato xxx/xx
CBU	Valor numérico de 22 dígitos
Cód. país, provincia, partido, localidad	Valor computado
Moneda	Formato de monedas generales
DNI, CUIL, CUIT	Valor numérico. Verificar su validez. Controlar 11 dígitos para el CUIL/CUIT.

Tabla 2. Niveles de criticidad primer y segundo marco

Nivel	Acción
Válido	Pasar a inspección del siguiente marco.
Modificable en Analítica	Realizar correcciones básicas o enriquecerlas y pasar a inspección del siguiente marco.
Modificable Usuario	Devolver al dueño del dato para su corrección o descarte.
Inválido	Informar al usuario que el dato es inválido y descartarlo. Queda a criterio de la organización su destino.

En la Tabla 3, se mencionan los indicadores y las tareas a realizar al momento de analizar y evaluar los datos. El indicador *Utilidad/Oportunidad/Actualidad* analiza si para lo que fueron tomados los datos, éstos resultan de utilidad desde el momento en que se insertaron en el repositorio o, desde la última vez que se procesaron o usaron.

El indicador *Correctitud* evalúa la cercanía de los valores de los datos a un conjunto de valores definidos, dentro de un dominio sintáctica y semánticamente correctos. Por ejemplo, un nombre que dijera “GERJO” se podría analizar y comparar con otro dato o información y determinar que significa “JORGE”.

Tabla3. Dimensión Disponibilidad

DISPONIBILIDAD	
Indicador	Significado
Utilidad/ Oportunidad/ Actualidad	Describe y evalúa la característica temporal del dato.
Correctitud	Datos léxica, sintáctica y semánticamente correctos.

En la Tabla 4, el indicador *Credibilidad* verifica si los datos existen en el rango de valores conocidos o aceptables.

Tabla 4. Dimensión Usabilidad

USABILIDAD	
Indicador	Significado
Credibilidad	Grado en el que los datos tienen atributos que se consideran ciertos y creíbles en un contexto de uso específico. La credibilidad incluye el concepto de autenticidad (la veracidad de los orígenes de datos, atribuciones, compromisos).

En la Tabla 5, se plantean cuatro indicadores. El indicador *Exactitud*, debe chequear si la representación de datos (o valor) refleja bien el estado real de la información de origen y no causa ambigüedad.

El indicador *Consistencia*, verifica si después de procesar los datos, sus conceptos, dominios de valor y

formatos son coincidentes y válidos como antes de procesarlos.

El indicador *Compleitud*, evalúa si el dato es necesario para su posterior análisis en el proceso de negocio. Por último, el indicador *Integridad*, evalúa si el dato fue manipulado para beneficios personales u organizacionales afectando su integridad. Además, controla si el dato se corresponde a dominios codificados y existentes.

Tabla 5. Dimensión Confiabilidad

CONFIABILIDAD	
Indicador	Significado
Exactitud	Los datos proporcionados son precisos, la representación de datos refleja el estado real de la información de origen y su representación no genera ambigüedades.
Consistencia	Consistencia luego del procesamiento, análisis de formato y dominios.
Compleitud	Porcentaje en que la información representa al mundo real.
Integridad	Verificar el cumplimiento de referencias informacionales, dominios codificados y chequear la manipulación de los datos relacionada con la completitud y exactitud. - El formato de los datos es claro y cumple los criterios. - Los datos son consistentes con la integridad estructural. - Los datos son consistentes con la integridad del contenido.

En la Tabla 6, el indicador *Conveniencia* corrobora si los datos de muestreo son provechosos en el contexto que se evalúan y si se relacionan en algún punto con el propósito para lo que se los tomaron.

Tabla 6. Dimensión Pertinencia

PERTINENCIA	
Indicador	Significado
Conveniencia	- Los datos recogidos no coinciden completamente con el tema, pero exponen un aspecto. - La mayoría de los conjuntos de datos recuperados están dentro del tema de recuperación que los usuarios necesitan. - El tópico de la información proporciona coincidencias con el tema de recuperación de los usuarios.

En la Tabla 7, se exponen dos indicadores. El primero evalúa el *Contexto Estadístico*, verificando si los datos analizados están cercanos a la media de los valores o por debajo/encima de ella.

Por otro lado, se debe analizar el *Contexto Externo* a través de indicadores y otros aspectos que se consideren, para así determinar si lo que resultó de la analítica es consistente con lo obtenido externamente. En resumen, chequear la consistencia de la información por medio de referencias externas.

Tabla 7. Dimensión Referencia Externa

REFERENCIA EXTERNA	
Contexto Estadístico	Se utiliza la estadística para evaluar la información e identificar valores atípicos y erróneos.
Contexto Externo	Consistencia de la información por medio de referencia externa.

Los niveles de criticidad de datos de este marco son los mismos que los que definidos en la Tabla 2. Cabe mencionar que esta tabla, utilizada para evaluar el nivel de criticidad de los dos primeros marcos, es de tipo general y las acciones definidas pueden incluir mejoras a los elementos propuestos para el marco de aseguramiento de calidad de datos para analítica.

Asimismo, se considera que si bien estas acciones son orientativas, cada organización puede tomar otras acordes a sus políticas y tratamiento de la información que considere necesarias.

El análisis realizado dentro de cada una de las dimensiones de este marco permite chequear, controlar y otorgar un nuevo nivel de criticidad al grupo de datos evaluados, permitiendo avanzar al siguiente marco, denominado de negocio.

4.4. Marco de Negocio

En la actualidad, el procesamiento de datos en analítica es realizado principalmente para que sus resultados otorguen una oportunidad de mejora para el negocio [12].

No obstante, a veces puede suceder que si bien los datos son físicamente correctos o tiene características en su contexto válidas, no sean útiles para el negocio o el propósito para el que fueron tomados otorgando resultados falsos.

Por consiguiente, se debe tener en cuenta ciertos aspectos prescriptivos del negocio, utilizando especialmente la inferencia del analista.

El análisis prescriptivo se centra en informar acerca de lo que debería suceder buscando mejorar el resultado esperado. En la Tabla 8, se definen dos indicadores a considerar para evaluar estos aspectos.

Una vez cumplido este último marco se debe calificar al dato dentro de un nivel de criticidad final de calidad para determinar con qué grupo de datos se trabajará en analítica y así esperar obtener los resultados deseados, es

decir, aquellos beneficiosos, efectivos y útiles para el negocio.

Tabla 8. Análisis de Marco de Negocio

ANALISIS PRESCRIPTIVO	
Indicador	Significado
Documentos: Conformación del documento	Si se está analizando un documento específico se debe conocer su formato y verificar la completitud de sus partes. Por ejemplo, cabecera y detalle de una boleta o factura.
Geo posiciones: Verificación Dato/información según referencia temática del mismo	De acuerdo a la meta de la información tomada, se analiza si el geo posicionamiento, validado física y contextualmente, es aceptable a nivel del negocio sutil del mismo. Por ejemplo, si se analiza la posición de un polígono, verificar que la geo posición no indique un río.

En la Tabla 9, se indica el nivel de criticidad final esperado luego de cumplimentar todas los marcos evaluados.

Tabla 9. Niveles de criticidad final

Nivel	Acción
Válido	Pasa al grupo de análisis en analítica.
Modificable en Analítica	Se aplican correcciones básicas o se enriquecen y pasan al grupo de análisis de analítica.
Inválido	Se descarta la información.

5. Resultados obtenidos y esperados

Como respuesta al interrogante planteado, se han definido los elementos necesarios para definir un marco general de aseguramiento de calidad de datos en entornos de analítica. Estos elementos incluyen tres secciones denominadas “marco”, con dimensiones, indicadores y acciones/tareas a realizar en cada una de ellas, para lograr alcanzar buenos niveles de calidad para posteriormente procesarlos analíticamente.

A partir de la definición de estos elementos, se continúa trabajando en la metodología, los procesos a incluir en cada marco especificado y las métricas de evaluación de cada uno de los indicadores, para obtener finalmente una propuesta integral que garantice datos de calidad aceptable en entornos de analítica.

Con el marco integral de aseguramiento de calidad propuesto, se busca examinar, clasificar y filtrar datos para que luego se conviertan en activos auténticos y

válidos, que permitan identificar nuevas oportunidades [17].

Asimismo, se pretende que este modelo marco contribuya a disminuir la cantidad de datos no oportunos o veraces, para que sólo permanezcan aquellos que cumplan el fin para el que fueron tomados, ya que trabajar con datos de mala calidad puede llevar a resultados erróneos con consecuencias en el futuro económico del negocio.

En la Figura 2, se muestra la arquitectura final que se desearía alcanzar una vez desarrollado el marco de aseguramiento de calidad para analítica. Asimismo, se espera que la evolución de este marco metodológico lo convierta en un nuevo estándar utilizable por las organizaciones, para que las mismas certifiquen la calidad de datos en sus repositorios para analítica.

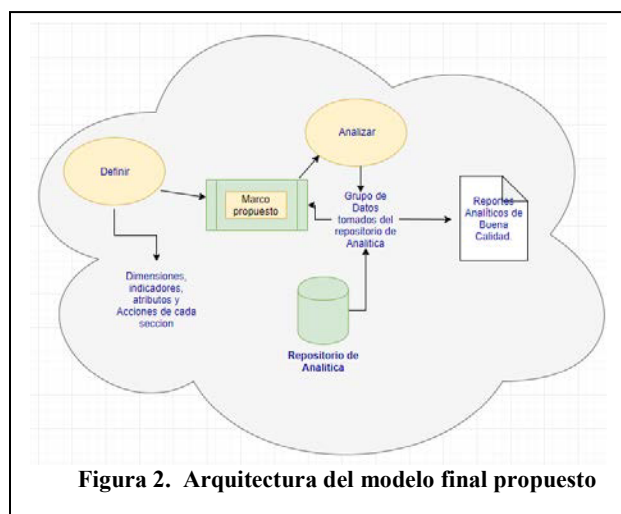


Figura 2. Arquitectura del modelo final propuesto

6. Conclusiones

La calidad de los datos para una organización es un tema que toma relevancia año tras año. Sin embargo, todavía falta definir estándares que aseguren su calidad, especialmente si se desea evaluar los repositorios de datos para analítica, aspecto del que no se tienen referencias que aborden el tema. Obtener activos de información de buena calidad es el punto más importante para concluir en informes con tendencias de mercado y ventajas corporativas reales.

La definición de los elementos y la especificación de una metodología con procesos como propuesta de marco de aseguramiento de la calidad busca mejorar la toma de decisiones de las organizaciones que necesitan analizar grandes volúmenes de datos, por lo que se puede decir que implementar un estándar para la mejora de la calidad de los datos facilitará encontrar información consolidada y correcta [25, 26].

Entre las contribuciones y beneficios que se espera obtener con esta propuesta se destaca:

- La aplicación de un modelo marco que permita a los encargados de analítica de datos encontrar información con verdadero sustento que conlleve a buenos resultados para el negocio.
- El resultado de su implementación en las áreas analíticas permitirá trabajar con una base sólida de datos y sustentable.
- Una implantación satisfactoria de la calidad de los datos a nivel organizacional permitirá maximizar el retorno de las inversiones en BI (Business Intelligence o Inteligencia de Negocios) logrando una ventaja competitiva y liderazgo en el mercado [27].
- Contar con datos precisos y oportunos para el correcto funcionamiento de la organización. Sin ellos, los sistemas de BI pueden suponer una amenaza que debilite la capacidad para evaluar el estado real de la misma y poder competir de forma eficaz [27].

Es importante destacar que, como toda metodología con procesos, puede ser adaptado y ajustado por cada organización a su propio sistema de trabajo.

Como líneas futuras de investigación en este ámbito se plantea continuar ampliando la especificación de los procesos y metodología aplicados a cada marco propuesto, definir un conjunto de métricas para evaluar los indicadores del marco de contexto e incluir una cuarta etapa de monetización (valor) del dato. Dichos agregados deben seguir las bases de cada una de ellas y tener la capacidad para ser aplicados al grupo de datos que se encuentra en proceso de análisis.

7. Referencias

- [1]Baskarada, S., Koronios, A. (2017). Unicorn data scientist: the rarest of breeds. Program 51(1): 65-74 (2017)
- [2] Carretero, A., Velthuis, M. (2018). Importancia de la Calidad de los Datos en la Transformación Digital. RUIDERAE: Revista de Unidades de Información. Número 13 (1er semestre 2018). ISSN 2254-7177
- [3] Hernández González, Arnulfo N. (2014). "Características de Calidad de Datos de los Almacenes de Datos". España: ResearchGate.
- [4] Benavides, G.R. and Sáenz R.B. (2001). "Aseguramiento de la calidad analítica y norma ISO 17 025 en laboratorios clínicos y Químicos", Revista Costarricense de Ciencias Médicas. ISSN 0253-2948, vol. 22 n.1-2, San José, Jun. 2001.
- [5]Acevedo. (2019). "La importancia de la calidad de los Datos para Big Data". Universidad Sergio Arboleda.

- [6] Power Data. (2014), "El valor de la Gestión de Datos". Categoría: Data Quality. Publicación de Power Data hacia Febrero 2014. España.
- [7] Davenport T., Dyché, J. (2013). Big Data in Big Companies. International Institute for Analytics.
- [8] Tecnología para los Negocios - Big Data: Los efectos positivos de la analítica de datos en la empresa. <https://ticnegocios.camaravalencia.com/servicios/tendencias/los-efectos-positivos-de-la-analitica-de-datos-en-la-empresa/> Último Acceso Septiembre 2019
- [9] Alarcos ResearchGroup. (2015). "MAMD: Towards a Data Improvement Model Based on ISO8000-6X and ISO/IEC 33000".
- [10] Wang, R. (1998) "A Product Perspective on Total data Quality Management". Communications of the ACM. Vol. 41, Issue 2, pp. 58-65. 1998.
- [11] ISO, ISO/IEC 25012 (2009) Software engineering - Software product Quality requirements and evaluation (SQuaRE) - data quality model, Ginebra: International Organization for Standardization, 2009.
- [12] Logicalis (2019). El proceso y los retos para obtener datos fiables para tu analytics. E-book. Barcelona, España.
- [13] Yordan Andreu Alvarez (2015). "Análisis de la calidad de datos en fuentes de la suite ABCD". Universidad Central "Marta Abreu" de las villas. Santa Clara
- [14] Creswell, J. (2002). "Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research". Prentice Hall. ISBN 10: 01-3613-550-1.
- [15] Riveros, H. y Rosas, L. (1985). "El Método Científico Aplicado a las Ciencias Experimentales". México: Editorial Trillas. ISBN 96-8243-893-4.
- [16] Argimón J. (2004). "Métodos de Investigación Clínica y Epidemiológica". Elsevier España, S.A. ISBN 9788481747096.
- [17] Manterola, Carlos; Astudillo, Paula; Arias, Esteban; Claros, Nataniel & Grupo MINCIR (2011). "Metodología e Investigación en Cirugía, Revisiones sistemáticas de la literatura. Qué se debe saber acerca de ellas". España: Elsevier, Cirugía Española.
- [18] Basili, V. (1993). "The Experimental Paradigm in Software Engineering". An Experimental Software Engineering Issues: Critical Assessment and Future Directions (Ed. Rombach, H., Basili, V., Selby, R.). Lecture Notes in Computer Science, Vol. 706. ISBN 978-3-540-57092-9.
- [19] ISO 9000, (2015). <https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:es>. Último Acceso Septiembre 2019
- [20] ISO 9126. } <https://www.issco.unige.ch/en/research/projects/ewg96/node13.htm>. Último Acceso Septiembre 2019
- [21] ISO 15939 (2007). <https://www.sis.se/api/document/preview/909004/>. Último Acceso Septiembre 2019
- [22] ISO 14598-5 (1998). <https://www.iso.org/obp/ui/#iso:std:iso-iec:14598:-5:ed-1:v1:en>. Último Acceso Septiembre 2019.
- [23] ISO 8000. <http://iso8000.es/normas-iso-8000>. Último Acceso Septiembre 2019
- [24] ISO 33000. <http://www.iso33000.es/index.php/familia-iso-33000>. Último Acceso Septiembre 2019
- [25] Rivera Tavera, L. (2013). "Herramienta para el mantenimiento, integridad y aseguramiento de la calidad de datos". Universidad Piloto de Colombia. Bogotá D.C.
- [26] Analítica Negocios (2019). "Analítica de Datos". Panamá: Analítica de Negocios.
- [27] Vega, Bianca (2012). "Hacia un Modelo de Madurez para la Gestión de Calidad de Datos en Inteligencia de Negocios". Departamento de Ingeniería de Sistemas y Computación Universidad Católica del Norte. Antofagasta, Chile.